Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 40 (2007) 605–608

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Guest Editorial

Intelligent data analysis in biomedicine

1. Introduction

Data are ubiquitous in biomedicine. In the domain of clinical care, information systems longitudinally capture and store hundreds of data elements for each patient at each outpatient encounter and inpatient hospitalization. Sources of these data include diagnostic as well as therapeutic services. Patient data management systems in operating theaters and intensive care units even collect megabytes of monitoring data during a single day. New three-dimensional digital imaging techniques have created vast arrays of patient data, the potential of which is largely yet to be explored. In molecular biology, the emergence of high-throughput DNA sequencing technology has created new, rich sources of genomic information.

Data have also bourgeoned in clinical and basic research facilities and laboratories. The increasing use of advanced assay methods in genomic, proteomic, and metabolomic studies and high-performance database architectures for research data capture, storage, and analysis has resulted in the availability of unprecedented amounts of data. This development affects virtually every region of the clinical and basic research domain, and stresses the importance of translational research. This is reflected by new funding initiatives such as the National Institutes of Health's Roadmap and Clinical Translational Science Award (CTSA) programs.

This increase in the amount and availability of data in clinical care and research settings carries a concomitant burden: the effective, efficient, and accurate analysis of these data. Traditional statistical methods, such as linear, logistic, and Cox regression modeling, have served us well over the years, and continue to do so. They constitute the foundational analytical toolkit for those working in clinical care, research, and healthcare administration, even as new statistical tools continue to be developed and find their way into mainstream use in these domains [1]. However, statistical methods rely on specialized knowledge and skills of analysts, and require them to carry out a large number of manual operations. As a result, a "data gap" is caused by the widening gulf between the amount of biomedical data and the number and skills of those who would analyze such data [2,3]. In addition, the sheer size of today's biomedical data collections often causes problems for traditional statistical methods. These matters inform the purpose and focus of this special issue of the *Journal of Biomedical Informatics* on intelligent data analysis in biomedicine.

Intelligent data analysis (IDA) can be defined as the use of specialized statistical, pattern recognition, machine learning, data abstraction, and visualization tools for analysis of data and discovery of mechanisms that created the data [2]. Such data are typically complex, meaning that they are characterized by many records, many variables, subtle interactions between variables, or a combination of all three. In addition, variable-value domains may be expressed in ways that depart from the traditional nominal–ordinal–interval-ratio hierarchy, such as log scale or model-based representations. Finally, and especially with the evolution of translational research, biomedical data are becoming increasingly heterogeneous, with clinical observation and measurement data merged with genomic or proteomic data in the same record. All of these contribute to a complex data landscape that is often difficult if not impossible to analyze without the use of IDA-based tools. The 15 papers in this issue provide a strong and "cutting-edge" representation of these tools and their applications.

Before we review the themes that emerge from the papers in this issue, we must first address the relationships of IDA with its close relatives, *knowledge discovery in databases* and *data mining*. Knowledge discovery in databases (KDD) is best thought of as a *process* where the data are used for hypothesis *generation* rather than hypothesis *testing* (as in traditional statistical analysis). It is conducted within a lifecycle framework that includes data cleaning, data preparation, model building, model evaluation, visualization, and reporting [4–6]. The process of KDD is structured around this framework to provide an overarching analytic methodology and to avoid a "fishing expedition". Certainly, hypothesis testing is often used in KDD, but is generally restricted to exploration of the data or the comparative evaluation of two or more analytic tools. Thus, point estimates of chi-square values and t-statistics, along with p-values and confidence intervals, will often be reported in a KDD analysis, but not as definitive indicators of association or implied causation. These would be left to the application of statistical methods in a more traditional analysis, which may have been informed by a KDD analysis.

1532-0464/$ - see front matter
doi:10.1016/j.jbi.2007.10.001

Beyond its hypothesis-generating focus, another characteristic that distinguishes KDD from traditional statistical analysis is the use of different methods from the disciplines of pattern recognition and signal detection, statistics, and machine learning. Although sometimes the terms are used interchangeably, KDD is not the same as *data mining*, which is the *application* of specialized tools to assist with the process of KDD. More precisely, data mining is the application of specialized software tools to the analysis of large data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the owner [7]. IDA is closely related to data mining, but uses prior domain knowledge to guide the analysis, often in an interactive and iterative fashion. Using prior knowledge not only increases the efficiency of the knowledge discovery process, but also helps to avoid reporting trivial results to the end-user. IDA uses many of the same methods used in traditional analysis, but also includes statistical classifiers such as support vector machines and hidden Markov models, probabilistic methods such as Bayesian classifiers and belief networks, nature-inspired methods such as neural networks, genetic algorithms, and artificial immune systems, as well as tree-building and rule-discovery methods. Many of these are represented in this special issue.

In the biomedical field, IDA emerged in the late 1980s from a quest for objective knowledge sources that could replace the traditional 'expert' in the development of rule-based systems [3]. Subjectively defined rules may mirror the psychological biases that are common to everyday human reasoning. Furthermore, expert rule bases were often found to be inconsistent and incomplete despite the fact the individual rules appeared to be correct. For these reasons, the learning of rules from data that were gathered in the daily practice of experts and stored systematically in databases became appealing as a more reliable source of knowledge. Within this approach, experts obtained a different role than before, providing background knowledge for focusing and guiding the learning process. With the increasing volumes of data that are today being captured in clinical practice and biomedical research, the emphasis of IDA has shifted from knowledge-based systems engineering to the more general task of interpreting data in an "intelligent" way. What has remained unchanged, however, is the central role of knowledge in the data analysis process: IDA uses knowledge and data to discover new knowledge.

By way of introduction to the papers in this issue, we have considered two important dimensions of IDA: methods and applications. Each of these dimensions in turn has its own set of themes that emerge from the papers, and we discuss these here.

2. Methods of IDA

2.1. Dimensionality reduction and feature subset selection

There are numerous methodologic approaches to IDA that are reflected in the themes within which the papers in this issue coalesce. The KDD life cycle provides a conceptual framework for considering these approaches. The first step in this life cycle is data cleaning and preparation, which frequently involves the reduction of the dimensional complexity of a dataset. When we talk about the complexity of biomedical data, we are often actually thinking about their dimensionality and, in turn, ways to reduce it so that we can discover meaningful (i.e., understandable) knowledge. *Feature subset selection* is an essential component of this reduction process and is represented in this issue by several papers.

Panteris et al. [8] describe a new methodology for feature selection using metabolic pathways and prior domain knowledge to create "pathway signatures" as an approach to dimensionality reduction in microarray data. Mahata and Mahata [9] report on a method for selecting subsets of predictive genes from thousands of candidates ranked by minimum probability of classification error. The novelty in their work is that these probabilities are computed by non-parametric density estimation. Nayak and De [10] also investigate reducing the complexity of signaling pathway information, but through the use of a new modularization algorithm.

2.2. Data exploration

Exploring a dataset is an important prelude to building models for classification and prediction; it is the component of the KDD life cycle where most knowledge discovery actually takes place. There is a critical need for the development and application of software tools that facilitate accurate knowledge discovery through user-directed exploration. These tools provide users the opportunity to interact with them, so that they have some sense of input into the exploration process; they stand in contradistinction to purely automated exploration tools that leave users out of that process. The incorporation of user-defined knowledge and user interaction in data exploration is a hallmark of the IDA-based approach to KDD. One family of such exploration tools relies on visualization of relationships between features, between values of features, and between features and classes. Demsar et al.'s paper [11] is an example of the kind of work that is being done in data visualization that goes beyond traditional graphical methods such as histograms and scatterplots. Demsar et al. present a tool, FreeViz, that facilitates the human comprehension of highly complex classification data, by visual exploration of multidimensional feature interactions. These are often unexpected, but common in biomedical data. A second paper (Lessmann et al. [12]) describes a method for visualizing complex histopathological image features, informed by human expert knowledge in order to extract information from the image that is both meaningful and relevant. Finally, Dinu et al. [13] report on a new tool, Pathway/SNP, which is designed to help users explore the association between biomolecular pathways and disease by integrating domain knowledge with statistical and data mining algorithms.

Clinical data are intrinsically temporal, as clinical events take place at a particular moment in, or over some period of time [14]. It is thus appropriate that a number of papers in this issue investigate the development and evaluation of methods and tools for time-series or temporal data exploration. One particular challenge in this context is the extraction of temporal patterns that are both relevant for the application domain at hand and meaningful for users. Guyet et al. [15] address this challenge by using a multi-agent collaborative learning framework that incorporates human input to analyze multivariate time-series data. Agent-based autonomy, adaptability, and emergence provide the foundation for the framework, and this work represents one of the first investigations of multi-agent technology in IDA. The paper by Sacchi et al. [16] also focuses on time-series data analysis but from the perspective of temporal abstraction. Sacchi et al. use data to create Precedence Temporal Networks, a novel, graph-based approach to visualization of temporal precedence relations. Wallstrom and Hogan [17] consider the discovery of aggregates from time series data, but using an unsupervised learning approach that incorporates Markov Chain Monte Carlo simulation to estimate the parameters in complex Bayesian clustering models.

2.3. Classification and prediction

While data exploration and visualization are essential elements of IDA, the development of useful and accurate models for classification and prediction represents the potential for applying IDA to real-world problems in basic medical and clinical research and ultimately clinical practice. Several papers in this issue report on the development and evaluation of IDA tools for classification and prediction. The two papers by Verduijn et al. [18,19] describe a novel predictive modeling method that employs a Bayesian network composed of a collection of local supervised learning models, which are in turn recursively learned from the data. An interaction layer ensures that clinicians can use the tool in real-world settings to determine patient prognosis. Wang [20] also applies probabilistic graphical models for prediction, but now in the genomic domain for the discovery of interactions between genes and transcription factors.

Support vector machines (SVMs) have been used with some frequency in genomics, but less so in clinical domains. Matheny et al. [21] investigate four methods for optimizing parameters for two different kernel SVMs when applied to a clinical problem, the prediction of mortality in percutaneous coronary interventions. They show that SVMs can be very sensitive to the selected optimization method. A novel time-series approach to clinical prediction is investigated by Toma et al. [22]. This hybrid method incorporates the discovery of frequent temporal patterns within traditional logistic regression models. The resulting models are then used for predicting patient mortality in critical care.

3. Application domains of IDA

As evidenced by the papers in this issue, IDA techniques can be applied to a broad base of application domains in biomedicine. For instance, clinical applications range from critical care (the papers by Verduijn; Toma; Guyet) to cardiology (Metheny), pathology (Lessmann et al.), and biosurveillance (Wallstrom and Hogan). Reflecting the success of modern DNA sequencing technology, the issue also includes a number of basic science applications of IDA: the discovery of gene expression regulation mechanisms (Sacchi et al.; Wang), analysis of biochemical signaling pathways in the cell (Panteris; Dinu; Nayak), and discovery of molecular diagnostic markers (Mahata and Mahata). This issue is completed with an excellent review of knowledge-based data mining methods for gene expression data by Bellazzi and Zupan [23], which complements an earlier review on predictive data mining in clinical medicine by the same authors [24].

4. The future of IDA in biomedicine

Knowledge discovery in databases has recovered from a ''trough of disappointment'' that followed the initial hype during the dot-com era of the late 1990s, and has evolved from an activity that was carried out primarily by academic researchers to one that is becoming part of the mainstream business process [25]. Yet there are many challenges that remain, especially in the biomedical field where large-scale electronic gathering and storage of data have just started. With improvements in the standards for coding and exchanging biomedical information and knowledge, collections of molecular and clinical data will become not only larger but especially more *comprehensive* as they link heterogeneous information on a single individual and even across entire populations of individuals. In addition, repositories of background knowledge, as in the form of web ontologies for example, will continue to grow. These sources present growing challenges for IDA researchers to develop methods that will help to further our understanding of the complex relationships between genetic disposition, psychosocial development, lifestyle, medical care, and health. We hope that the papers included in this issue will stimulate researchers working in IDA and related disciplines to explore and develop new methods and tools and find new applications for them, while encouraging those who are new to IDA to consider it as a fruitful field for research and practice in biomedicine.

References

[1] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.

[2] Grossman RL, Page C, Greeley R, et al., editors. Data mining for scientific and engineering applications. Dordrech: Kluwer; 2001.

[3] Lavrac N, Keravnou E, Zupan B. Intelligent data analysis in medicine. In: Kent A et al., editors. Encyclopedia of computer science and technology, Vol. 42. New York: Dekker; 2000. p. 113–57.
[4] Frawley W, Piatetsky-Shapiro G, Matheus C. Knowledge discovery in databases: an overview. In: Piatetsky-Shapiro G, Frawley W, editors. Knowledge discovery in databases. AAAI/MIT Press; 1991.
[5] Han J, Kamber M. Data mining: concepts and techniques. San Francisco: Morgan Kauffmann; 2000.
[6] Brachman RJ, Anand T. The process of knowledge discovery in databases: a human-centered approach. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining. Cambridge, MA: MIT Press; 1996.
[7] Hand D, Mannila H, Smyth P. Principles of data mining. Cambridge: The MIT Press; 2001.
[8] Panteris E, Swift S, Payne A, Liu X. Mining pathway signatures from microarray data and relevant biological knowledge. J Biomed Inform 2007;40:698–706.
[9] Mahata P, Mahata K. Selecting differentially expressed genes using minimum probability of classification error. J Biomed Inform 2007;40:775–86.
[10] Nayak L, De RK. An algorithm for modularization of MAPK and Calcium signaling pathways: comparative analysis among different species. J Biomed Inform 2007;40:726–49.
[11] Demsar J, Leban G, Zupan B. FreeViz. An intelligent multivariate visualization approach to explorative analysis of biomedical data. J Biomed Inform 2007;40:661–71.
[12] Lessmann B, Nattkemper TW, Hans VH, Degenhard A. A method for linking computed image features to histological semantics in neuropathology. J Biomed Inform 2007;40:631–41.
[13] Dinu V, Zhao H, Miller PL. Integrating domain knowledge with statistical and data mining methods for high density genomic SNP disease association analysis. J Biomed Inform 2007;40:750–60.
[14] Shahar Y. Dimensions of time in illness: an objective view. Ann Intern Med 2000;132(1):45–53.
[15] Guyet T, Garbay C, Dojat M. Knowledge construction from time series data using a collaborative exploration approach. J Biomed Inform 2007;40:672–87.
[16] Sacchi L, Larizza C, Magni P, Bellazzi R. Precedence temporal networks to represent temporal relationships in gene expression data. J Biomed Inform 2007;40:761–74.
[17] Wallstrom G, Hogan WR. Unsupervised clustering of over-the-counter healthcare products into product categories. J Biomed Inform 2007;40:642–8.
[18] Verduijn M, Peek N, Rosseel PJM, de Jonge E, de Mol BAJM. Prognostic Bayesian networks. I. Rationale, learning procedure, and clinical use. J Biomed Inform 2007;40:609–18.
[19] Verduijn M, Rosseel PJM, Peek N, de Jonge E, de Mol BAJM. Prognostic Bayesian networks. II. An application in the domain of cardiac surgery. J Biomed Inform 2007;40:619–30.
[20] Wang J. A new framework for identifying combinatorial regulation of transcription factors: a case study of the yeast cell cycle. J Biomed Inform 2007;40:707–25.
[21] Matheny ME, Resnic FS, Arora N, Ohno-Machado L. Effects of SVM parameter optimization on discrimination and calibration for post-procedural PCI mortality. J Biomed Inform 2007;40:688–97.
[22] Toma T, Abu-Hanna A, Bosman R-J. Discovery and inclusion of SOFA score episodes in mortality prediction. J Biomed Inform 2007;40:649–60.
[23] Bellazzi R, Zupan B. Towards knowledge-based gene expression data mining. J Biomed Inform 2007;40:787–802.
[24] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2006. doi:10.1016/j.ijmedinf.2006.11.006.
[25] Piatetsky-Shapiro G. Data mining and knowledge discovery 1996–2005: overcoming the hype and moving from "university" to "business" and "analytics". Data Min Knowl Disc 2007;15:99–105.

Guest Editors
John H. Holmes
Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA
E-mail address: jhholmes@mail.med.upenn.edu

Niels Peek
Academic Medical Center, Universiteit van Amsterdam, Department of Medical Informatics, Meibergdreef 9, 1105AZ Amsterdam, The Netherlands
E-mail address: n.b.peek@amc.uva.nl

Available online 12 October 2007